複数の辞書の定義文の照合に基づく同義表現の自動獲得

村田　真樹[†]　金丸　敏幸[††]　井佐原　均[†]

近年，言い換え表現の自動獲得の研究が重要視されつつある．本稿では，複数の辞書を用意して，それらにおける同じ項目の定義文を照合することにより，言い換え表現の一種である同義表現を抽出することを試みた．また，同義表現を抽出するための新しい尺度を提案し，その尺度で抽出データをソートした結果の精度は，一般によく行なわれる頻度だけでソートする方法による結果よりも高いことを確認した．この尺度は，他の同義表現の抽出の研究にも利用できる有用なものである．提案手法では，同義表現のみを正解とするとき，上位500個で0.748，ランダムに抽出した500個で0.220の抽出精度であった．また，誤りの多くのものは包含関係や類義関係にある表現であり，それらも正解と判断する場合は，上位500個で0.954，ランダムに抽出した500個で0.722の抽出精度であった．

キーワード:　言い換え，抽出，辞書，定義文，同義語

Automatic Paraphrase Acquisition Based on Matching of Definition Sentences in Plural Dictionaries

Masaki Murata[†] , Toshiyuki Kanamaru[††] and Hitoshi Isahara[†]

Studies on paraphrasing are important in various research topics such as sentence generation, summarization, and question-answering. Extracting automatic paraphrases by matching definitions of the same word in two dictionaries is described. A new method for extracting these paraphrases is also described. Higher precision was obtained than with the conventional method of using frequency. Our method can be applied to other studies on paraphrase extraction. The method obtained the precision rate of 0.748 in the top 500 data and that of 0.222 in the 500 data that were extracted randomly, when a synonym only was judged as a correct answer. It obtained the precision rate of 0.954 in the top 500 data and that of 0.722 in the 500 data that were extracted randomly, when a hypernym and a similar expression were also judged as correct answers.

KeyWords:　*Paraphrase, Extraction, Dictionary, Definition Sentence, Synonym*

1　はじめに

言い換えに関する研究 (佐藤 1999; Murata and Isahara 2001; 乾 2002; 村田，井佐原 2004) は平易文生成，要約，質問応答 (村田，内山，井佐原 2000; 村田 2003) と多岐の分野において重要なものであり，近年，その重要性は多くの研究者の認めるところとなっている．また，これと同時に，言い換え表現の自動獲得の研究も重要視されつつある．本稿では言い換え表現の一

† 情報通信研究機構, National Institute of Information and Communications Technology
†† 京都大学, Kyoto University

種である同義表現を自動獲得する研究について述べる．本稿では，複数の辞書を用意して，それらにおける同じ項目の定義文を照合することで，同義表現を抽出する．

例えば,「あべこべ」という語の定義文を考えてみる．大辞林では，

「順序・位置などの関係がさかさまに入れかわっていること。」

となっており，岩波国語辞典では，

「順序・位置・関係がひっくり返っていること。」

となっている．これらの定義文は同じ「あべこべ」という語の定義文であるため，同義な内容を記述した文であり同義なテキスト対と見ることができる．これを照合すれば，

「さかさまに入れかわっている」

⇕

「ひっくり返っている」

といった同義な表現対が得られる．本稿の手法は大雑把には以上のとおりで，このように同義な内容を記述する複数の辞書の定義文を照合することで同義表現を獲得するのである．

本研究の価値をあらかじめ整理すると以下のようになる．

- 同義なテキスト対から同義表現を抽出する研究はいくつかあるが，複数の辞書の定義文を同義なテキスト対として，そこから同義表現を獲得する先行研究はない．本稿は，複数の辞書の定義文からどのくらいの同義表現を抽出できるかの目安を与えるものとなる．
- 本稿では，同義表現の抽出に役に立つ，新しい尺度を提案する．本稿の実験で，この尺度がいくつかの比較手法よりも有効であることを確認する．この尺度は，他の同義表現の抽出の研究にも利用できる有用なものである．

2　複数の辞書の照合に基づく同義表現の抽出方法

本研究では，複数の辞書を用意して，それらにおける同じ項目の定義文を照合することにより，同義表現を抽出することを試みる．この辞書としては，岩波国語辞典と大辞林を使用した．同義テキスト対としては，二つの辞書の各見出し語の定義文同士を組にすればよいが，場合によっては一つの見出し語が複数の項目をもっている場合がある．これの対処法として，本稿ではそれぞれの定義文が，岩波国語辞典と大辞林とで一対一に対応すると仮定して，照合の度合いが良いもの同士，定義文を結び付けることにした．

まず照合のとりかたであるが，これは各定義文を JUMAN(黒橋，長尾 1998) を使って形態素列に分解する[1]．各行に形態素が来るようにして UNIX の diff コマンドを使って，一致，不一

1 定義文を JUMAN を使って形態素列に分割したが，定義文の解析において JUMAN の性能が特に下がるということはなかった．また，本研究ではこの形態素列への分解の処理においては JUMAN を使用しただけであり，辞書を追加したり後処理をするなどの他の処理はしていない．

表 1　辞書定義文の照合結果の例

照合の度合い	見出し語	定義文の diff の結果
0.69	あいこ	＜互いに、＞≦たがいに≧勝ち負けのないこと
0.29	あえか	＜はかなげな＞≦たよりない≧さま
0.17	あえか	＜美しくかよわげな＞≦かよわく、なよなよした≧さま
0.20	あからさま	＜急な＞≦包み隠さないで、はっきり表す≧さま
1.00	あさって	あすの次の日
1.00	あしらい	もてなし
1.00	あたふた	あわてふためくさま
0.17	あたふた	＜数量が非常に多い＞≦あわただしく動作を急ぐ≧さま
0.40	あたら	＜惜しい＞≦もったいない≧ことに
0.18	あたら	＜もったいなく＞≦おしく≧も
0.22	あっさり	＜濃かったり、くどかったり、しつこかったりせず、＞さっぱり＜としたさま＞
0.67	あっぷあっぷ	水におぼれかけて＜、もがいている＞≦苦しむ≧さま
0.54	あとり	スズメよりやや＜大形で頭と背面は黒色＞≦大形≧
0.35	あとり	＜日本へは＞秋に≦シベリア地方から日本に≧渡来＜し、全土で越冬＞する
1.00	あべこべ	反対
0.41	あべこべ	順序・位置＜などの＞≦・≧関係が＜さかさまに入れかわって＞≦ひっくり返って≧いる＜・＞こと

致箇所を検出する (村田，井佐原 2002)．照合の度合いを計る式としては，以下のものを用いた．

$$\text{照合の度合い} = \frac{\text{一致文字数} \times 2}{\text{全文字数}} \tag{1}$$

ここで，一致文字数は，diff の結果一致部分と判断された部分の文字数を意味し，全文字数は，diff に与えた岩波国語辞典と大辞林の双方の定義文を合わせた文字数を意味する．この式は，0 から 1 の値をとり，一致部分が大きいほど大きな値を持つものとなっている．

実際に上記の照合を行なった．照合は 57,643 個の定義文の対で行なうことができた．辞書定義文の照合結果の例を表 1 に示す．表中で “＜”,“＞” で囲まれた部分は，大辞林にだけ出現したものを，また，“≦”,“≧” で囲まれた部分は，岩波国語辞典にだけ出現したものを意味する．

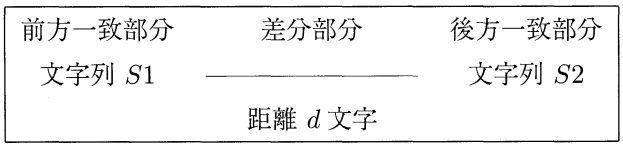

図 1　差分の出現模式図

表をみると,「互いに」と「たがいに」や,「惜しい」と「もったいない」や，さきほどの「あべこべ」の「さかさまに入れかわって」と「ひっくり返って」といった同義表現が得られていることがわかる．しかし,「急な」と「包み隠さないで、はっきり表す」や,「数量が非常に多い」と「あわただしく動作を急ぐ」といった誤った対応のものも見受けられ，この結果をそのまま用いるのは精度が悪そうである．

そこで，次に，diff の結果から，ある程度良さそうな同義表現を抽出することを試みる．ここでは以下の二つの特徴を利用することにする．

- 珍しい (出現頻度の低い) 文字列に囲まれた差分部分ほど，同義表現としては確からしい．
- 複数箇所に出現した差分部分ほど，同義表現としては確からしい．

まず，一つ目の「珍しい文字列に囲まれた差分部分ほど，同義表現としては確からしい」という特徴の方を考える．ここでは，差分部分が図 1 のように，一致部分である文字列 $S1$,$S2$ に挟まれていて，$S1$ と $S2$ の間が d 文字だけ離れているとする．本稿では，この d としては，差分部分の長い方の文字数を採用する．このとき，$S2$ および $S1$ からみて，d 文字以内に図 1 のように $S1$ および $S2$ が現れる確率を，$P(S1)$，$P(S2)$ とすると，$P(S1)$，$P(S2)$ は d 文字以内の各箇所で $S1$ または $S2$ が出現しない事象の余事象の確率となり，以下のように表される．

$$P(S1) = 1 - (1 - \frac{\text{文字列 } S1 \text{ の出現数}}{\text{文字総数}})^{(d+1)} \tag{2}$$

$$P(S2) = 1 - (1 - \frac{\text{文字列 } S2 \text{ の出現数}}{\text{文字総数}})^{(d+1)} \tag{3}$$

このときの差分部分が確からしい確率を $P(\text{差分}, S1, S2)$ とすると，$P(\text{差分}, S1, S2)$ は $S1$,$S2$ がともに図 1 のような形で現れにくい確率であると仮定すると，以下のようになる．($S1$ と $S2$ が独立であることを仮定している．)

$$P(\text{差分}, S1, S2) \simeq (1 - P(S1))(1 - P(S2)) \tag{4}$$

次に，二つ目の「複数箇所に出現した差分部分ほど，同義表現としては確からしい.」を考える．これは，複数箇所での確率をうまく組み合わせればよい．複数箇所のうち一か所でも正しければ，その差分部分は正しいものとして抽出できると考える．つまり，差分部分が正しい事象は，任意の $S1$,$S2$ に対して $S1$,$S2$ に囲まれる差分部分がすべて確からしくない場合の余事象なので，差分部分が確からしい確率を $P(\text{差分})$ とすると，それは以下の式で表される．(各差

分部分が独立であることを仮定している．)

$$P(\text{差分}) \simeq 1 - \prod_{S1,S2} (1 - P(\text{差分}, S1, S2)) \tag{5}$$

この値を尺度としてデータをソートし，この値が大きい差分ほど同義表現として確からしいと判断する．便宜上，この値でデータをソートする方法を村田法と呼ぶことにする．

3　比較手法

本節では，前節の村田法の有効性を確かめるために実験で用いる比較手法について述べる．

- 頻度法 — 差分部分の頻度に基づいてデータをソートする方法である．頻度の高いものほど，同義表現として確からしい差分部分と判断する．
- 加藤第一法 — これは加藤らの文献 (加藤，浦谷 1999) を参考にして作成した方法である．以下の条件式を満足する差分部分だけを使って頻度法に基づいて，同義表現として確からしい差分部分を判断する．

$$\frac{\text{「文字列 } S1 \text{ の文字数」} + \text{「文字列 } S2 \text{ の文字数」}}{d} > 1 \tag{6}$$

差分部分の長さよりも，差分部分の環境を構成する文字列 $S1$ と文字列 $S2$ の長さの和の方が長いほど，偶然生じた差分でなく対応の取れた差分部分と判断できる．そのために作成された方法である．

- 加藤第二法 — これは加藤第一法と村田法を組み合わせた方法である．式 (6) の条件式を満足する差分部分だけを使って村田法で用いる式 (5) に基づいて，同義表現として確からしい差分部分を判断する．

4　実験

まず，実際に村田法に基づいて抽出結果をソートしてみた．その結果を，表 2 に示す．表の「評価」は獲得された差分部分に対する評価を意味し，「◎」と「○」は，差分が同義と判断される文脈があることを意味し，「<」は，左の差分が右の差分に意味的に包含されると判断される文脈があることを意味し，「>」は，左の差分が右の差分を意味的に包含すると判断される文脈があることを意味する．「◎」は差分の両方ともになんらかの表現がある場合で「○」は差分の片側が空文字の場合である．ここでいう「文脈」とは，抽出される文章での文脈以外でも結果の判定者が思いついた文脈でもよく，判定者の思いついた文脈において同義や包含関係などが成立した場合もそのように判断される文脈があるとした．表の「*」は，判定者の思いついた文脈に書き換えたことを意味し，それ以外の文脈は抽出の際に用いられた前方一致部分と後方一致部

表 2　差分部分の抽出結果 (上位 50 個)

評価	$-\log(1-P)$	頻度	前方一致部分の例	差分部分		後方一致部分の例
○	4995	786	^心配がなく		、	のんびりしているさま$
○	3135	424	^たちが悪い		・	こと$
◎	2402	318	^また、その問題	・	や	質問$
◎	2267	301	^等級や段階	が	の	低いこと$
〈	1761	550	^目や口		など	を急に大きく開くさま$
○	1531	234	^楽器の金属		の	弦$
○	851.9	87	^ガイガー		ー	カウンター$
○	761.3	105	^外部と連絡・交渉		を	すること$
◎	706.3	89	を「にくい」と	いう	言う	類$
○	564.4	68	アナウンサー		」	の略$
○	527.1	133	七菜、二の膳		に	五菜、三の膳
○	376.0	72	^くろぐろ		と	している$
◎	375.1	51	竹の繊維を材料	と	に	して作った紙$
○	352.5	117	乗車券などを、		その	当日より
○	343.2	62	^別のもので		は	ないこと$
◎	299.7	39	^有理数	・	と	無理数の総称$
○	285.5	47	^書物をのせて読む		ための	台$
○	232.6	44	^収入が支出より		も	多いこと$
◎	223.5	27	他の役所	に	へ	文書で通知すること$
◎	217.4	32	^実権	が	を	伴わない
◎	213.9	60	^印刻に	使う	用いる	小刀$
◎	212.5	30	^全体の重量	・	または	容量$
◎	209.7	31	^会計年度など	で、	のを	一年を半分
◎	203.1	33	^森林	の	を	保護
◎	200.7	26	^密輸出	と	やが	密輸入$
○	190.9	32	^また、興ざめ		する	$
◎	177.5	52	^その人の利益となる	こと	事	を主張して助ける
◎	169.3	54	ひじを曲げた	とき	時	、二の腕に
○	166.5	35	^警視正の下		で	、警部の上$
◎	164.0	22	^濁音	で	に	発音する$
◎	162.7	22	^正しい解答	または	やまた、	解釈$
○	158.4	921	、			食用
◎	130.3	55	手加減をしないで	もの	物	を言うさま$
◎	128.1	29	^一割の	一〇	十	分の一$
◎	126.9	28	、徴兵検査	で	のよく	甲種合格
〉	125.6	26	^ぐっすりと			眠ること$
◎	123.9	16	^家の中央	にある	の	部屋$
〉	122.7	36	を受けず、自分の力		だけ	で生活を
◎	116.6	23	^船がドックに	はいる	入る	こと$
◎	115.0	16	^襲いかかって	くる	来る	こと$
○	115.0	29	少しずつにじみ出る		ような	さま$
◎	111.4	14	皇子・内親王	、	・	天皇の弟などの総称$
○	110.0	23	、会合に出席		したり	すること$
◎	108.4	17	味を消すため	、	に	別の物を
◎	106.0	22	^製造	する	のつ	方法$
○	105.8	178	数が一		つ	であること$
◎	104.6	22	^大正末期	に	の	流行$
◎	101.5	23	して銑鉄を	つくる	作る	こと$
◎	101.0	12	^踊りを職業と	している	する	女性$
○	99.17	33	^口や胃		の中	に入れたものを

表 3　獲得された同義表現の例

つつ	ながら
一六	十六
哺乳動物	哺乳類
中途	途中
業	職
である	となる
なる	変わる
隔たり	差
つく	到着する
で作った	の
家畜	牛馬など
がうまい	に巧みな
大事に	大切に
伝える	伝達する
ために	目的で
はずれている	合わない
食う	食べる
減少する	少なくなる

分を意味する.「^」は定義文中での文頭を,「$」は文末を意味する.

「・」「など」「の」を片一方では省略するなどの規則の他，主格の際の「が」と「の」や,「一〇」と「十」や,「または」と「や」や,「使う」と「用いる」などの同義な言い換え表現も獲得されていることがわかる.

ここにあげたもの以外に得られた良さそうな同義表現を表 3 にあげておく．すでにある同義語辞書にも登録されていそうな単語レベルの同義語だけでなく,「がうまい」と「に巧みな」のようなフレーズレベルのものから,「つつ」と「ながら」のような機能的な同義語なども獲得できている.

次に比較手法と数量的な比較を行なった．この結果を表 4 と表 6 にあげる．ここでは，差分が同義と判断される文脈があれば，その差分は正解と判断した．表 4 は，それぞれの手法でソートした結果での上位数個の差分での精度である．表 6 には，村田法と加藤法での抽出精度と抽出数を整理している．加藤法は加藤第一法と加藤第二法の総称である．表 6 の抽出精度はランダムに取り出した 500 個での精度で，抽出総数はそれぞれの方法で得られた差分の総数である．予測抽出数は抽出総数と抽出精度を掛け合わせたもので，それぞれの手法により抽出できそうな同義表現の総数を意味する．加藤法では式 (6) で差分を足切りするので，抽出総数は村田法よりも減る．また，頻度法は足切りをしないので，抽出総数は村田法と同じ結果となる.

表 4 の結果を見て欲しい．上位での精度は式 (5) を用いる村田法と加藤第二法が良く，われわれの提案する式 (5) でソートする方法が有効であることがわかる．頻度法と加藤第一法の比較では，ソート自体は同じ頻度を用いるのに精度は加藤第一法の方が良い．加藤法での式 (6) に

表 4　抽出精度 (上位 500 個)

	村田法	頻度法	加藤第一法	加藤第二法
上位 50 個	0.940 (47/ 50)	0.580 (29/ 50)	0.680 (34/ 50)	0.900 (45/ 50)
上位 100 個	0.890 (89/100)	0.560 (56/100)	0.620 (62/100)	0.860 (86/100)
上位 200 個	0.795 (159/200)	0.580 (116/200)	0.645 (129/200)	0.810 (162/200)
上位 300 個	0.777 (233/300)	0.583 (175/300)	0.657 (197/300)	0.767 (230/300)
上位 400 個	0.760 (304/400)	0.590 (236/400)	0.642 (257/400)	0.760 (304/400)
上位 500 個	0.748 (374/500)	0.588 (294/500)	0.616 (308/500)	0.736 (368/500)

表 5　抽出精度 (上位 500 個．差分の片側が空文字の場合を除く)

	村田法	頻度法	加藤第一法	加藤第二法
上位 50 個	0.960 (48/ 50)	0.880 (44/ 50)	0.920 (46/ 50)	0.980 (49/ 50)
上位 100 個	0.960 (96/100)	0.900 (90/100)	0.930 (93/100)	0.960 (96/100)
上位 200 個	0.945 (189/200)	0.910 (182/200)	0.905 (181/200)	0.950 (190/200)
上位 300 個	0.930 (279/300)	0.903 (271/300)	0.907 (272/300)	0.927 (278/300)
上位 400 個	0.915 (366/400)	0.907 (363/400)	0.895 (358/400)	0.907 (363/400)
上位 500 個	0.904 (452/500)	0.910 (455/500)	0.878 (439/500)	0.876 (438/500)

表 6　抽出精度と抽出数

	村田法	加藤法
抽出精度	0.220 (110/500)	0.400 (200/500)
抽出総数	67851	17104
予測抽出数	14927	6841

よる差分の足切りは精度において効果があることがわかる．

次に表 5 の結果を見てみよう．この表は差分の片側が空文字の場合を除いた結果である．差分の片側が空文字の場合は，片一方で単に詳しく述べているだけの場合や，対応づけの誤りである場合もあり，同義表現としてはふさわしくない対が多い．この結果は，表 4 の結果に比べて格段に良くなっている．また，この結果でも村田法は他の手法に比べて比較的良い精度をあげている．

最後に表 6 を見てみよう．ランダムに 500 個を取り出したときの精度は，村田法は 0.22 と低く，また，加藤法は 0.40 と高い．これは先に述べたのと同じように加藤法では式 (6) によりあ

表 7　関係の種類

	同義関係	包含関係	類義関係	関係なし
ランダム	0.220 (110/500)	0.454 (227/500)	0.048 (24/500)	0.278 (139/500)
上位 50 個	0.940 (47/ 50)	0.060 (3/ 50)	0.000 (0/ 50)	0.000 (0/ 50)
上位 100 個	0.890 (89/100)	0.100 (10/100)	0.000 (0/100)	0.010 (1/100)
上位 200 個	0.795 (159/200)	0.165 (33/200)	0.000 (0/200)	0.040 (8/200)
上位 300 個	0.777 (233/300)	0.187 (56/300)	0.000 (0/300)	0.037 (11/300)
上位 400 個	0.760 (304/400)	0.205 (82/400)	0.003 (1/400)	0.033 (13/400)
上位 500 個	0.748 (374/500)	0.202 (101/500)	0.004 (2/500)	0.046 (23/500)

表 8　関係の種類 (差分の片側が空文字の場合を除く)

	同義関係	包含関係	類義関係	関係なし
ランダム	0.313 (106/339)	0.274 (93/339)	0.071 (24/339)	0.342 (116/339)
上位 50 個	0.960 (48/ 50)	0.020 (1/ 50)	0.000 (0/ 50)	0.020 (1/ 50)
上位 100 個	0.960 (96/100)	0.010 (1/100)	0.000 (0/100)	0.030 (3/100)
上位 200 個	0.945 (189/200)	0.030 (6/200)	0.000 (0/200)	0.025 (5/200)
上位 294 個	0.932 (274/294)	0.031 (9/294)	0.007 (2/294)	0.031 (9/294)

らかじめ不確かな差分を削除しているためで，このため加藤法は精度が良い．しかし，抽出総数は加藤法は格段に減ってしまい，予測抽出数は村田法の方が加藤法よりもはるかに高い．加藤法では多くの同義表現を取りこぼす問題があることがわかる．

次に抽出総数も多く，また，上位での精度も高かった村田法の抽出結果についてより詳細な調査を行なった．その結果を表 7 と表 8 に示す．表 7 のランダムは，ランダムに取り出した 500 個での結果を意味し，上位 X 個は上位 X 個での結果を意味する．この調査では得られた差分が，同義表現と判断される文脈があるか，差分同士で意味的に包含関係にあると判断される文脈があるか，類義表現と判断される文脈があるか，それ以外かで集計し整理した．表 8 は，表 7 の調査で差分の片側が空文字の場合を除いて集計した結果である．包含関係，類義関係にある表現と判断した表現の例を表 9 と表 10 に示しておく．この調査は，同義表現でないとされた誤りの多くが，包含関係にある表現や類義表現であったために行なったものである．

同義関係だけでなく，包含関係や類義関係を正解と判断するとかなり高い精度となることがわかる．全データでは，0.722 $(= 1 - 0.278)$ 程度の精度を持つことになる．差分の片側が空文字の場合を除いた場合で，0.658 $(= 1 - 0.342)$ 程度の精度を持つことになる．包含関係や類義

表 9　包含関係にある表現の例 (左の表現が右の表現を意味的に包含する)

前方一致部分の例	差分部分		後方一致部分の例
^目や口	など		を急に大きく開くさま$
^二匹	以上		の蚕が
方向を示す		大	方針$
^餅を		たくさん	食べたあとの、もたれた感じの
^多角形の	すべての	各	頂点が
方針や見込み	などの	が	立たないこと$
、	細長い	ひも状の長い	舌で
花が	全部	いっせいに	咲く$
^穴を	埋める	埋めて平らにする	こと$
主君	・親	や父	など
^左大臣の	異称	唐名	$

表 10　類義関係にある表現の例

前方一致部分の例	差分部分		後方一致部分の例
二三度	二七	二六	分の緯線$
^	ほり	池	や
霧のように	一面に広がった層状の	低くただよう	雲$
^	悪かったと	過失や人に迷惑を掛けたことを	あやまること$
^	大きな	非常な	利益$
^銅を	とる	含む	鉱石$
海上の	国防	防衛・攻撃	を
^都を	追われて	立ち去って、	地方
授業	などを休む	等が休みになる	日$

関係も正解と判断する場合は，差分の片側が空文字の場合を除くとかえって精度が下がるようである．これは一見不思議なことだが，よく考えるとそれほど不思議なことではない．包含関係にある表現を示した表 9 の例を見て欲しい．差分の片側が空文字の場合は，空文字でない側の差分に「など」「以上」のように意味を広げる効果を持つ表現が来る場合は空文字でない側が空文字の側の差分を意味的に包含する関係となり，空文字でない側の差分になんらかの意味的な限定をする表現が挿入されると空文字でない側の差分の方が意味が狭くなり空文字でない側が空文字の側の差分に意味的に包含される関係となる．このため，差分の片側が空文字の場合は包含関係にある表現であることが多く，包含関係や類義関係を正解と判断する場合は，差分の片側が空文字の場合も含めた場合の方が精度が高くなるのである．

また，包含関係や類義関係を正解と判断する場合は，上位での精度も非常に高いものとなる．例えば，全データでの上位 500 個だと，包含関係や類義関係を正解と判断すると，精度は 0.954 $(= 1 - 0.046)$ となる．

以上の結果をまとめると，以下のようになる．

- 式 (5) を用いる村田法は，上位での精度が高い．また，一般によく行なわれる頻度でソー

トするだけよりも，精度が高いことも確認された．村田法を利用することで，効率よく精度の高い同義表現の知識を獲得することができる．

- 抽出総数は，村田法の方が加藤法よりも多い．多くの同義表現を抽出する場合には，加藤法の式 (6) での差分の候補の除去はしない方がよい．
- 同義表現の抽出としては，差分の片側が空文字の場合を除いた方が精度が高い．しかし，包含関係や類義関係も正解と判断する場合は，差分の片側が空文字の場合も含めた方が精度が高い．
- 抽出総数も多く，また，上位での精度も高かった村田法の抽出結果は，同義表現のみを正解とするとき，上位 500 個で 0.748，ランダムに抽出した 500 個で 0.220 の精度で，包含関係や類義関係も正解と判断する場合は，上位 500 個で 0.954，ランダムに抽出した 500 個で 0.722 の精度であった．

5　関連研究

本節では関連研究について述べる．

まず，本稿と同様に同義なテキスト対を照合することによって，同義表現を抽出する研究について説明する．この種の研究に用いられる同義なテキスト対には，以下のようなものがある．それぞれについて説明する．

- 同じ原文からの複数の翻訳を利用する

  同じ原文から作成された翻訳を複数集めると，その翻訳同士は同義なテキスト対となる．このテキスト対を照合することで同義表現を獲得するのである (今村，秋葉，隅田 2001; Barzilay and McKeown 2001; 下畑，渡辺，隅田，松本 2003)．今村ら (今村他 2001) はこの方法で同等表現を抽出し，60％程度の精度で正しい言い換え文の生成を行なっている．Barzilay ら (Barzilay and McKeown 2001) も同様の方法を利用して言い換え表現の候補を 9483 対抽出し，精度はほぼ意味的に等価な言い換え対を正解として 85～92％の精度であったとしている．また，20 回以上出現した 112 個の言い換え表現では，WordNet において 35％が同義表現で 32％が上位下位関係であったと報告している．下畑ら (下畑他 2003) は抽出した同義表現を実際に機械翻訳の研究に役立てる研究もしており，抽出した同義表現の利用により翻訳可能な文を 8％向上させ，正しい翻訳を出力する割合も 2.5％向上させたと報告している．
- 同内容の記事対を利用する

  複数の新聞社の記事を収集し，同じ事柄を記述している記事群を抽出する．この同じ事柄を記述している記事群が同義なテキスト対となるのである．このテキスト対を照合することで同義表現を獲得するのである (関根 2001)．関根 (関根 2001) は，固有表現抽出の技術を使い，固有表現は対応づけのキーとしてテキスト対を照合する工夫をしている．

この手法は，同じ事柄を記述している記事群を探すところから，処理をする必要がある手間の多い手法ではあるが，新聞データはたくさんあるので，うまくいくと多くの同義表現を抽出できる可能性がある．関根の実験では1日の新聞記事から8つの言い換え表現を抽出しそのうちの4つが正しい言い換え表現であったとしている．

- 関連のある話し言葉と書き言葉のデータを利用する
  例えば，同内容の講演発表とその予稿を利用するのである．学会などでの発表を文字に書き起こしたデータと，その発表と同時に出す予稿の論文を利用して同義表現を抽出する研究として文献 (村田，井佐原 2001b, 2001a; Murata and Isahara 2002a) がある．発表とそれの元となった論文は，同内容のことを同一の著者が言った，また，書いたものなので，これらも同義なテキスト対と見なせるのである．このテキスト対を照合することでも同義表現を抽出することができる．論文の講演発表に限らず，ある発表とその発表の元になった書き言葉のテキストの対も同様に，同義なテキスト対と見なせるので，それらからも同義表現を抽出することができる．この論文では定量的な評価はなされていないが，実際に抽出された話し言葉と書き言葉の対が示されている．
- 同じ文書中の同内容の部分を利用する
  例えば，ある論文の要約の部分と，その論文全体は，要約がその論文の中身の要約であるので，文章の長さはかなり違うが内容は同じであるので，同義なテキストとみなせるのである．また，ある特許の請求項の節の内容と，その特許の実施例の節の内容も，同じ内容が記述されているので，同様に同義なテキストとみなせるのである．実際に，文献 (Murata and Isahara 2002b) では特許の請求項と実施例の対を利用して同義表現の抽出を試みている．この論文でも定量的な評価はなされていないが，実際に抽出された表現の対がいくつか示されている．
- 要約前のテキストと要約後のテキストを利用する
  要約の研究においては要約前のテキストと要約後のテキストを用意して，それを比較することで要約に関する同義表現を抽出する場合がある．このとき，要約前のテキストと要約後のテキストは同内容であるので，それらは同義なテキスト対とみなせるのである．このテキスト対からも同義表現を抽出することができる (加藤，浦谷 1999)．この論文での言い換え表現の評価では，上位100個のものはすべて要約用の言い換えとしては妥当なものであったとされている．同義表現としての評価はなされていない．

以上のように，同義なテキスト対を照合することで同義表現を抽出する研究には多くのものがある．ところで，本稿で提案している式 (5) で抽出結果をソートする村田法は，これらの研究にも利用できるものであり，村田法の適用範囲は広い．

次に以上の方法以外によって同義な表現を抽出する先行研究について述べる．これには，共起語が類似している単語同士を同義語とする研究がある (下村，福島 1993; Lin 1998; 山本 2002;

Lin, Zhao, Qin, and Zhou 2003)．共起語が類似していればその単語同士も類似している可能性が高く，類義語の抽出 (Hindle 1990) にはよく使われる方法だが，この方法を同義語の抽出にも使うのである．しかし，この方法では，同義語以外に反義語や類義語を多く抽出してしまうことが知られており，同義語の抽出には種々の工夫が必要な方法となっている．下村ら (下村，福島 1993) は類似度で取り出した上位 5208 対から 178 対の同義語を取り出したとしている．山本 (山本 2002) は類似度で抽出する手法に加えて種々のヒューリスティックを利用することで，66 %の精度で 1117 個の言い換え可能な表現と 114 個の双方向言い換え可能な表現を抽出したとしている．Lin ら (Lin et al. 2003) は類似度で抽出する手法に加えて，同義表現対が出現することのないパターンを利用して同義表現対かどうかの判断をすることにより，80 個の同義表現を用いた実験で適合率 86 %，再現率 95 % の精度でその同義表現を抽出できたとしている．

6　おわりに

言い換えに関する研究 (佐藤 1999; 乾 2002; 村田，井佐原 2004) は平易文生成，要約，質問応答 (村田他 2000; 村田 2003) と多岐の分野において重要なものであり，近年，その重要性は多くの研究者の認めるところとなっている．また，これと同時に，言い換え表現の自動獲得の研究も重要視されつつある．この状況を背景として踏まえ，本稿では，複数の辞書を用意して，それらにおける同じ項目の定義文を照合することにより，言い換え表現の一種である同義表現を抽出することを試みた．

その結果，本稿で新しく提案する式 (5) を用いる手法は，抽出データをソートした結果での上位での精度が高く，また，一般によく行なわれる頻度だけでソートする方法よりも，精度が高いことも確認された．この式 (5) を用いる手法は，他の同義表現の抽出の研究にも利用できる有用なものである．また，本稿の手法は比較手法に用いた加藤法よりも抽出総数が多い見込みを得ており，約 15,000 の同義表現を抽出できる見込みを得ている．同義なテキスト対から同義表現を抽出する研究はいくつかあるが，複数の辞書の定義文を同義なテキスト対として，そこから同義表現を獲得する先行研究はない．本稿は，複数の辞書の定義文からどのくらいの同義表現を抽出できるかの目安を与えるものとなっている．

また，同義表現の抽出としては，差分の片側が空文字の場合を除いた方が精度が高いが，包含関係や類義関係も正解と判断する場合は，差分の片側が空文字の場合も含めた方が精度が高いことがわかった．提案手法では，同義表現のみを正解とするとき，上位 500 個で 0.748，ランダムに抽出した 500 個で 0.220 の精度で，包含関係や類義関係も正解と判断する場合は，上位 500 個で 0.954，ランダムに抽出した 500 個で 0.722 の精度であった．

参考文献

Barzilay, R. and McKeown, K. R. (2001). "Extracting Paraphrases from a Parallel Corpus." In *39th Annual Meeting of the Association of the Computational Linguistics*, pp. 50–57.

Hindle, D. (1990). "Noun Classification from Predicate–Argument Structures." In *28th Annual Meeting of the Association for Computational Linguistics*, pp. 268–275.

今村賢治，秋葉泰弘，隅田英一郎 (2001). "階層的句アライメントを用いた日本語翻訳文の換言." 言語処理学会第 7 回年次大会ワークショップ論文集.

乾健太郎 (2002). "言語表現を言い換える技術." 言語処理学会第 8 回年次大会チュートリアル資料, pp. 1–21.

加藤直人，浦谷則好 (1999). "局所的要約知識の自動獲得手法." 言語処理学会誌, **6** (7).

黒橋禎夫，長尾真 (1998). 日本語形態素解析システム JUMAN 使用説明書 version 3.6. 京都大学大学院工学研究科.

Lin, D. (1998). "Automatic Retrieval and Clustering of Similar Words." In *COLING-ACL '98*, pp. 768–774.

Lin, D., Zhao, S., Qin, L., and Zhou, M. (2003). "Identifying Synonyms among Distributionally Similar Words." In *Proceedings of IJCAI-03*.

村田真樹，内山将夫，井佐原均 (2000). "類似度に基づく推論を用いた質問応答システム." 自然言語処理研究会 2000-NL-135, pp. 181–188.

村田真樹，井佐原均 (2001a). "同義テキストの照合に基づくパラフレーズに関する知識の自動獲得." 情報処理学会 自然言語処理研究会 2001-NL-142.

村田真樹，井佐原均 (2001b). "話し言葉と書き言葉の diff." ワークショップ「話し言葉の科学と工学」.

村田真樹，井佐原均 (2002). "diff を用いた言語処理 — 便利な差分検出ツール mdiff の利用—." 言語処理学会誌, **9** (2).

村田真樹 (2003). "質問応答システムの現状と展望." 電子情報通信学会学会誌, **86** (12), pp. 959–963.

村田真樹，井佐原均 (2004). "「言い換え」言い換えの統一的モデル — 尺度に基づく変形の利用—." 言語処理学会誌, **11** (5).

Murata, M. and Isahara, H. (2001). "Universal Model for Paraphrasing — Using Transformation Based on a Defined Criteria —." In *NLPRS'2001 Workshop on Automatic Paraphrasing: Theories and Applications.*

Murata, M. and Isahara, H. (2002a). "Automatic Extraction of Differences between Spoken and Written Languages, and Automatic Translation from the Written to the Spoken Language." In *LERC 2002.*

Murata, M. and Isahara, H. (2002b). "Using the Diff Command in Patent Documents." *Proceedings of the Third NTCIR Workshop (PATENT).*

佐藤理史 (1999). "論文表題を言い換える." 情報処理学会論文誌, **40** (7).

関根聡 (2001). "複数の新聞を使用した言い替え表現の自動抽出." 言語処理学会第 7 回年次大会ワークショップ論文集.

下畑光夫，渡辺太郎，隅田英一郎，松本裕治 (2003). "パラレルコーパスからの機械翻訳向け同義表現抽出." 情報処理学会論文誌, **44** (11).

下村秀樹，福島俊一 (1993). "共起類似性に基づく同義語の抽出." 情報処理学会第 47 回全国大会予稿集，1M-10, pp. 3-77–3-78.

山本和英 (2002). "テキストからの語彙的換言知識の獲得." 言語処理学会 第 8 回年次大会.

略歴

村田 真樹: 1993 年京都大学工学部卒業．1995 年同大学院修士課程修了．1997 年同大学院博士課程修了，博士（工学）．同年，京都大学にて日本学術振興会リサーチ・アソシエイト．1998 年郵政省通信総合研究所入所．現在，独立行政法人情報通信研究機構主任研究員．自然言語処理，機械翻訳，情報検索，質問応答システムの研究に従事．言語処理学会，情報処理学会，人工知能学会，電子情報通信学会，計量国語学会，ACL，各会員．

金丸 敏幸: 2001 年京都大学総合人間学部卒業．2003 年同大学院人間・環境学研究科修士課程修了．現在，同大学院博士課程在学中．認知意味論，自然言語処理の研究に従事．日本認知科学会，日本認知言語学会，日本語用論学会，各会員．

井佐原 均: 1978 年京都大学工学部電気工学第二学科卒業．1980 年同大学院修士課程修了．博士（工学）．同年通商産業省電子技術総合研究所入所．1995 年郵政省通信総合研究所．現在，独立行政法人情報通信研究機構けいはんな情報通信融合研究センター自然言語グループリーダー．自然言語処理，機械翻訳の研究に従事．言語処理学会，情報処理学会，人工知能学会，日本認知科学会，ACL，各会員．

(2004 年 1 月 8 日 受付)
(2004 年 4 月 30 日 再受付)
(2004 年 7 月 5 日 採録)